Comparing the end to end latency of an immersive collaborative environment and a video conference

Roberts, DJ, Duckworth, TW, Moore, CM, Wolff, R and O'Hare, JJ

http://dx.doi.org/10.1109/DS-RT.2009.43

Comparing the end to end latency of an immersive collaborative environment and a video conference

David Roberts, Toby Duckworth, Carl Moore, Robin Wolff and John O'Hare
Centre for Virtual Environments, University of Salford
Salford, UK
email: d.j.roberts@salford.ac.uk, t.w.duckworth@pgr.salford.ac.uk, c.m.moore@pgt.salford.ac.uk, rb.wolff@gmail.com, j.ohare@salford.ac.uk

Abstract - **Latency in a communication system can result in confusing a conversation through loss of causality as people exchange verbal and non-verbal nuances. This paper compares true end-to-end latencies across an immersive virtual environment and a video conference link using the same approach to measure both. Our approach is to measure end-to-end latency through filming the movements of a participant and their remote representation through synchronised cameras. We also compare contemporary and traditional immersive display and capture devices, whilst also measuring event latency taken from log files. We compare an immersive collaborative virtual environment to a video conference as both attempt to reproduce different aspects of the face-to- face meeting, the former favouring appearance and the latter attention. Results inform not only the designers of both approaches but also set the requirements for future developments for 3D video which has the potential to faithfully reproduce both appearance and attention.**

Keywords: tele-immersion, collaborative virtual environment, video conferencing, teleconferencing, virtual reality

I. Introduction

The aim was to gain an insight into how an Immersive Collaborative Virtual Environment (ICVE) measured up to a high quality video conference in terms of true end-to-end latency. The objectives were to compare the two by filming a persons actions and their remote representation through synchronised cameras, compare the impact of contemporary and traditional interfaces to the ICVE, and compare the impact of projector and monitor interfaces to the video conference. Measuring event latency within the ICVE assisted with second objective. Video conferencing and ICVE make an interesting comparison as both are methods to reproduce the face-to-face meeting that combine verbal communication with distinct aspects of non-verbal communication. Both can give the illusion of communicating both what someone looks like and what they are looking at. However, video conferencing faithfully reproduces appearance but not attention, while virtual environments faithfully reproduce attention but not appearance. Introducing latency between parties in a conversation can cause frustration and confusion. The tolerance for latency in verbal communication is 150 msecs. Verbal and non-verbal communication need to be synchronised but the level of synchronisation depends on the non-verbal resource, the visual representation and the application. For example lip sync and blinking in response to a question might require closer synchronisation than pointing to an object while talking about it. However, 150msecs gives us a safe target for end-to-end latency in multi-modal as well as just audio communication systems. While new immersive technology can certainly outperform old in terms of frame rate and image quality, there is some contention to how the end to end latency of a contemporary immersive and capture system compares to that of a traditional one. We therefore link contemporary and traditional style immersive systems, comparing true end-to-end and event latencies in both directions.

A. Related work

The addition of video to audio conferencing allows a range of non-verbal resources to improve participant's ability to show understanding, forecast response, enhance verbal descriptions, manage pauses and express attitudes [1], yet the 2D nature of video limits non-verbal communication, awareness and ability to point at and manipulate objects [2]. The ability to communicate eye gaze was compared across video conferencing and an immersive virtual environment [3], in a paper that concentrated on the spatial rather than temporal characteristics of the medium.

The Simulation Network Analysis Project (SNAP) investigated latencies within a flight simulator system [4] and while over a decade ago stands out as the most rigorous experiment of its kind. Timings were taken at the stick input at one simulator, a camera filming the display on a screen at the remote simulator and at numerous intermediate points. Special hardware was developed to time when information passed in and out of computer memory and network cards. GPS was used to synchronise measurements at two geographically remote sites. Our work may provide a more accurate true end-to-end latency measurement by using two hardware synchronised cameras but falls short of the rigour of internal measurement when compared to SNAP. As discussed at the panel session of the premier VR conference this year, latency has been largely overlooked in the majority of VR usability studies [5]. Latency in online computer games has been measured against network load [6]. Event traffic was reported when linking an immersive and desktop display in a Collaborative Virtual Environment [7]. The impact of update rate on event traffic has been reported for an eye-gaze enabled ICVE [8]. The end-to-end latency of a single user immersive virtual environment has

been measured by filming a person and the representation of their actions, both within one display by the same camera [9]. In a later study, a tracked pendulum replaced the user so that the delay could be taken from the phase difference of a sign wave [10].

Given the research uncovered in the above literature survey we believe the novelty of our paper lies in: comparing the end-to-end latency of video conferencing and ICVE in one experiment; testing true end-to-end latency of a visual communication system by syncing monitoring cameras through initial frame lock (genlock); measuring true end-to-end latency of a an ICVE; and measuring true end-to-end latency across an ICVE that links contemporary and traditional technology.

II. Approach

Exactly the same procedure was used to test end-to-end latency in both ICVE and video settings. We filmed a person repeatedly moving their arm up and down, both locally and as represented remotely through the medium, using synchronised cameras. Figure 1 shows the remote end of the ICVE setting, whereas Figure 2 shows both ends of the video conference setting.

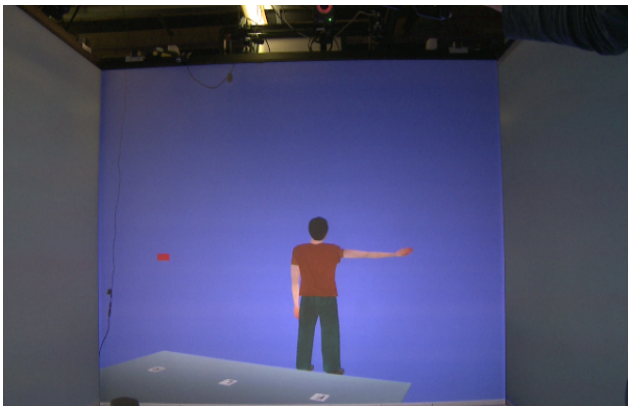


Figure 1. ICVE setting, showing avatar arm moving on the remote display.

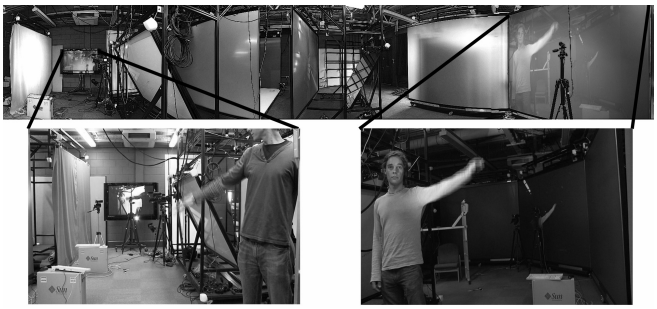


Figure 2. Video Conference setting, person moving arm and the video impression of this at the remote display.

We then identified the point at which upward arm movement turned downward on both videos and compared the frame count between the two. This was done many times, ignoring times when it was hard to see when the movement changed. The remote embodiment of the person being recorded differed between the mediums as the ICVE setting reproduced movement through a motion tracked avatar. Clock synchronisation is always an issue in such experiments. In order to measure the end-to-end latency, we synchronised two video cameras at the beginning of the experiment then took one to each end of the communication system. Synchronisation included genlock, ensuring synchronisation of the start of each frame. At the end of the experiment we brought the cameras back together and tested drift. No drift was found. In this synchronised mode the cameras ran at 50Hz. Unlike our previous experiments that had been between CAVE™-like systems in different parts of the country, we this time connected two CAVE™-like displays in different buildings on the same university campus and also connected two video conference nodes in the same room. This allowed camera drift to remain negligible as the cameras were taken to each end of the experiment. The video cameras used were Sony PMW-EX1.

In order to make a comparison between impact of interface and compute on timings we tested the event latency. Timestamped events were created at the local immersive display when the user moved within the scene using the joystick. They were then distributed to the remote site to update the position of the avatar. Latency was measured from the time the action showed an effect on the display of the local site (e.g. movement of the scene) until they were displayed at the remote site. These measurements were taken within the display layer of the ICVE, just before the scenegraph was updated and rendered on the screens. Thus, actual rendering time was not included. The timing measurements were taken from log files over the period of the experiment and the values are shown in milliseconds. As we were testing latency rather than human behaviour, experienced operators were used as test subjects as they understood the need for clear and regular movements and the operating constraints of the tracking system.

III. Experimental Conditions

Four conditions were provided through two technology approaches, each with two technology configurations, one at each end of the medium.

A. ICVE condition

The hardware environment consisted of cubic CAVE™-like displays in different building on the same university campus. The network between the two immersive displays has 1Gbit capacity. The virtual environment consisted of a room filled with chairs and small objects of various complexity. The entire scene, including avatars, consisted of around 13,500 shaded polygons. Our collaborative virtual environment EyeCVE [8] adopted the traditional replicated database approach of ICVEs. As someone moves within a virtual environment, the viewpoint into it should move with them. This is important not only for a feeling of presence, but for gauging where remote people are looking (Roberts et al, 2009). Critically, a perceptual lag in update of viewpoint can cause motion sickness. This is why decoupling

viewpoint scene rendering from network delay is necessary. Replicating the database and simulation at each site, while keeping sites synchronised through message passing allows viewpoint updates to be calculated locally. A typical ICVE system is shown in Figure 3.

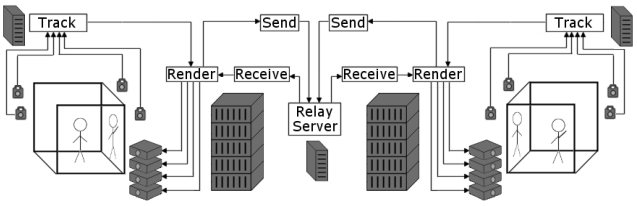


Figure 3. Diagram of typical ICVE system

EyeCVE differs from all other ICVEs in that it additionally communicates eye gaze. Various distributed modules make up the system. Each site runs an EyeCVE_Client and one server is used to download the world from and act as a tunnel through a firewall. All these modules communicate through UDP. Figure 4 shows the architecture of EyeCVE client. It consists of three layers: network, simulation and display. The network layer provides the communication tools to the client, transmits events between local and remote sites, and handles events based on selected consistency mechanisms, such as ordering and redundancy filtering. The simulation layer keeps a local database of objects inhabiting the virtual environment and updates their state based on user input from tracking or a controller, as well as the behaviour of specific objects. State changes are passed through the network layer, while update events coming in from remote sites are applied to the local database. The output layer of the client architecture is the display layer, which copies the object state from the simulation database to an object representation in a scenegraph and renders the scene. The EyeCVE client can run on a cluster of display nodes. This is implemented as a master/slave module within the display layer in the form of a distributed scenegraph. The master module copies all changes to its scenegraph to all slave nodes before rendering, see Figure 5. The slave nodes will then update their own copy of the scenegraph and render it. A slave module is implemented as a stand-alone display sever application, which runs on each display cluster node driving a specific screen or window.

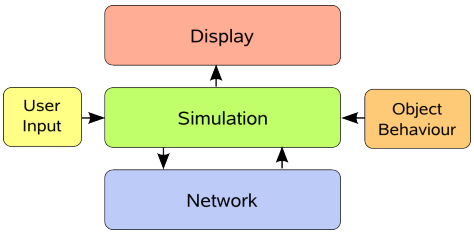


Figure 4. Architecture of EyeCVE

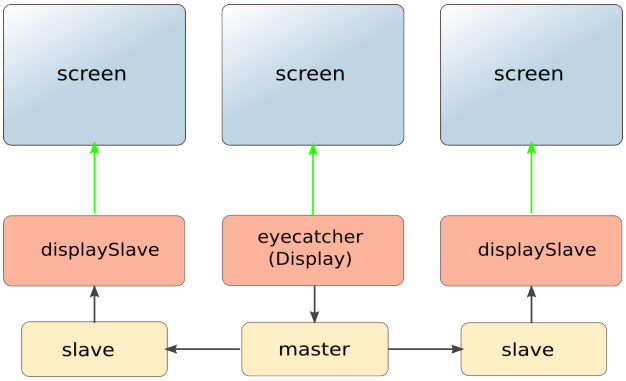


Figure 5. The display layer of a cluster client

In both of our immersive displays, surround projection was provided in a cube configuration with four walls and a floor. The walls were rear projected whereas the floor was projected from above.

1) Contemporary immersive display and capture: Each of these five surfaces displayed mono projection from a Christie S+3K Stereo DLP, running at a resolution of 1400x1050 at 102Hz. Active stereo was disabled for this experiment so that the results would reflect the passive stereo approach which seems to be becoming prevalent. The tracking system was an eight camera near infra-red optical tracker; Vicon MX-F40. Markers were placed on stereo glasses and the hand-held wand. The cluster implementation of EyeCVE ran across five graphics nodes connected across a 10Gbs, 250Mps Cisco 4900M switch. Each graphics node was a Sun Ultra 40 M2 Workstation, with quad-core 2.6Ghz AMD Opteron 64-bit processors, 8Gb memory and an Nvidia Quadro FX5600 graphics card. Each ran the rendering for a single viewpoint perspective, one for each projector. One additionally ran the EyeCVE client master node and another the log server. EyeCVE can run on this cluster under Windows or Linux, although active stereo does not work well under the former. As we wanted the results to be similar to the prevalent passive stereo Windows configuration we ran Windows with mono projection in this experiment.

2) Traditional immersive display and capture: The input and output of this display system are typical of those from when the building of CAVE installations was at its height, just over a decade ago. The compute system is about five years old and is typical of the upgrades to such installations around that time. The input system is a magnetic motion tracker. The projectors use Cathode Ray Tubes (CRT) and the computer is a shared memory multiprocessor device with multiple graphics cards. The screens were 3 metres wide and 2.5 high, active stereo projected from the rear by Barco 808s CRT projectors with a resolution of 1024x786 and a refresh rate of 100Hz (50Hz per eye). Head and hand tracking was done through a magnetic Ascension Flock of Birds tracker with an update rate of up to 140 Hz. Computation was provided by an SGI Prism system, fitted with six Intel Itanium2 1.4GHz CPUs, 6GB RAM, and four ATI FireGL graphic cards.

B. *Video Conference condition*

The cameras were Basler Pioneer megapixel (1004x1004) colour GigE. These ran at 48Fps. Unlike a conventional video conferencing set up, the cameras were networked and fed over IP directly to the displaying computer. The cameras were run in push mode so that frames were continually streamed to the receiving PC. The displays were set up at far ends of a large room each powered by a single computer. A single cable ran from each camera to the computer driving the remote display. For compute we used two Sun Ultra 40 M2 Workstations, each with dual-core 1.8Ghz Opteron 64-bit processors and 8Gb memory and an Nvidia Quadro FX5600 graphics card. Two display technologies were used allowing us to compare between plasma TV and Digital Light Projection (DLP). The large screen display was 250x200cm, and was rear DLP projected with a resolution of 1400x1050 and brightness 300cd/m2, from a Christie. This ran at 39 frames per second, as monitored by the AMCap video display software through vsync. The TV was full high definition 1080P (progressive scan) wide screen plasma colour monitor, made by LG. This ran at 33 frames per second, again as monitored by the AMCap video display software.

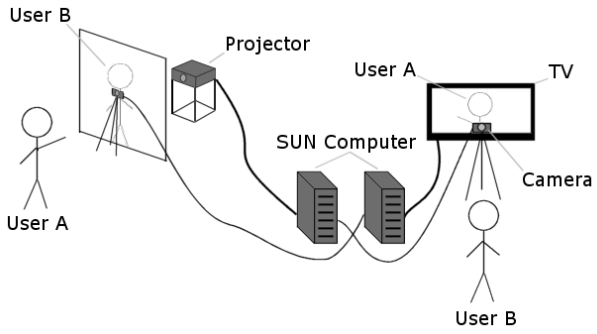
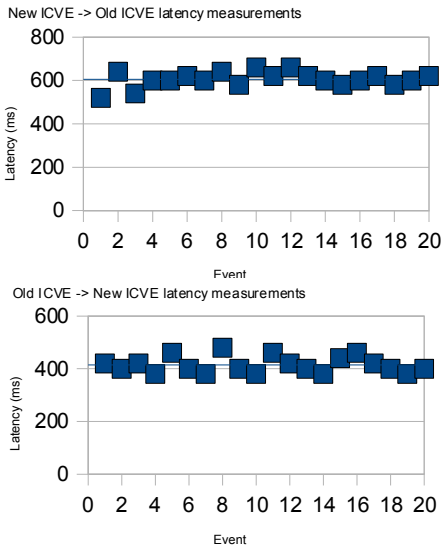


Figure 6. Video conference set up.

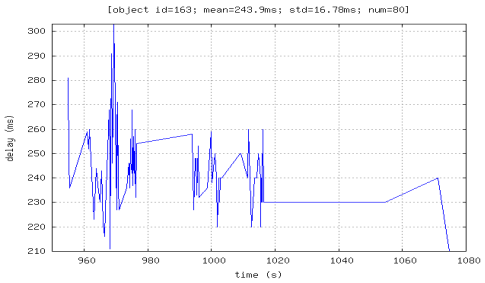
IV. RESULTS

The mean end-to-end delay from the contemporary to the traditional immersive display was 605msecs and that in the other direction was 414msecs. The data these measurements were taken from is shown in Figure 7 a and b respectively. The end-to-end latency of the video conference when viewed through the plasma screen had a mean of 120ms, and that viewed through the DLP 100ms. The event latencies of the ICVE are shown in figure 8. The mean values are of less value here as heavy variance is seen at the start of the session. This is because the ICVE system blocks when loading the model of the remote avatar from disk when users join the collaborative session. After the avatar is loaded, most of the events from contemporary to traditional fall between 220 and 260msecs and those in the other direction between 15 and 150msecs. The near zero and negative values in figure 7 b suggest a skew between the wall clocks. The directional bias is thus likely to be less than the above figures suggest.

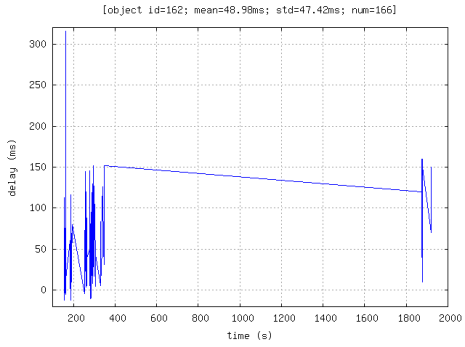


(b) Traditional to contemporary

Figure 7. The time it takes for the movement of a person to be replicated in the remote avatar: (a) from contemporary to traditional immersive display – mean of 605msecs; (b) in the other direction – mean 414msec.



(a) Contemporary to traditional



(b) Traditional to contemporary

Figure 8. Graph showing the latency in replicating enactment of an event: a) from contemporary to traditional immersive display – mean 250 b) from traditional to contemporary – mean 50.

V. Discussion

There are many factors that impact on latency in an ICVE. While processing has become faster, modern immersive displays often use slower projection and tracking technology than those built a decade ago. DLP projectors typically offer a much better image to cost ratio than CRT, are lower maintenance and have longer life span. However, most are slower than typical CRT projectors, and those that are not tend to cost more than CRT. This increased latency is somewhat offset by the move to passive stereo which runs at twice the frame rate of active stereo as the latter uses alternate frames for each eye. Although our DLP projectors were active stereo, we used them in mono so that the results would be representative of the faster passive approach. One of the worst problems in using immersive projection technology has traditionally been getting tangled in the wires of the tracking system. Now that vision processing is becoming faster, there has been a shift to optical tracking of markers placed on the subject. Optical technology also scales favourably to other approaches allowing full body tracking. Both these advantages are important to collaborative environments as the first stops confusion arising when one person gets tangled in wires the other can not see and the second gives the opportunity to communicate a wider set of non-verbal cues. Given that the latencies of projectors and tracking systems are published on specification sheets, we did not measure those used in this experiment. Instead we hoped to give a feel for how an old approach of shared memory, active stereo CRT projection and wired tracking measured up to a new approach of cluster compute, passive stereo DLP projection and wireless optical marker tracking. We also wanted to see what happened when the two were connected together.

Accuracy of synchronisation of monitoring in a distributed simulation is a crucial issue. Both the accuracy of synchronisation and granularity of measurement of the cameras used to record end-to-end latency was 20ms, which is the length of a frame. The synchronisation of the wall clock cannot be trusted to the degree of that of the cameras as there is no hardware sync. However the granularity of the wall clock is above that of the cameras. The wall clock can describe time to within a milisecond granularity, however, neither its reading or synchronisation is run on a dedicated processor through a real-time operating system. Furthermore, the wall clock synchronisation algorithm used does not account for directional bias in networking and processing. On this network there should have been no directional bias. We use this synchronisation approach as it is lighter weight than the more robust skew approach and in our previous experience, has performed favourably when run alongside computer simulation and graphics which are highly competitive for resources. However, several very low and even minus event latencies were measured but only in one direction. While two events may run through the various cycles of the system at different speeds, just as two cars may have different luck with traffic lights, this could only partly explain the low and not the negative results. This suggests that there was some clock skew.

In the ICVE condition both the true and internal event appear in the new technology in around two thirds of the time that their movements are seen by the other. What we can be reasonably sure of is that overall the event latency accounts for a total of a third of the true end-to-end latency in our system that combines old and new technology. If we do not account for clock skew in the event monitoring then the maximum event latency is almost double that in one direction to the other (240msec & 135msec), in stable conditions. The bias in event latency is in the same direction as that in end-to-end latency. The wall clock used for the monitoring was, however, skewed in the same direction, thus accounting for some of the event bias.

The VC configuration had an end-to-end latency within the tolerances of spoken conversation. However, this was an optimal set up in a single room and did not go through a single switch or router. The plasma screen increased the end-to-end latency by 20% above that when using a DLP projector. In the video conference we chose to send directly from the camera to the receiving computer via IP, rather than using a capture PC local to the camera as we believe that ultimately it is the best approach. It does require that the camera can address the receiving process.

It would be interesting to do a formal comparison with the industry standard for video conferencing, Access Grid. Access Grid specifies considerably lower resolution but an HD version is now available. While we have not undertaken formal experiments, our casual experiments and frequent use have observed latencies of over a second. The end-to-end latency of our ICVE system was about fivefold that of the high performance HD video conferencing. However the former went through both level two and three network switches whereas the latter was a direct cable. The ICVE uses far less bandwidth than the HD video. In a previous experiment over a link of several hundred miles, using traditional interfaces at both ends, event latencies were around 150msecs [8].

VI. Conclusion

By filming a participant and their remote representation, with cameras that stay in sync after being synchronised, we were able to gain reliable end-to-end latency measurements for both an Immersive Virtual Environment and a video conference. We were able to build an HD video conference prototype that had significantly less end-to-end latency than an ICVE prototype. However, the ICVE was running across campus whereas the video conference was running within one room. Both approaches demonstrated lower latency than we have become used to in frequent Access Grid meetings. Event logging provided an internal view of the timings within the ICVE. The interface of the ICVE, including graphics rendering, accounted for around two thirds of the total latency, whereas a third was from event propagation and simulation. When a traditional immersive interface was connected through our ICVE to a contemporary one, the person in the older system sees the movements of the other quicker. The bias in event delay is in the same direction.

However, the wall clock used for measuring events seems to have been skewed exaggerating the latter.

The video system fell within the 150ms latency tolerance of a multi-modal communication system that includes voice, whereas in the faster direction the ICVE exceeded this limit by almost three fold.

Acknowledgements

This work was funded by the EPSRC eye-catching project under the grant EP/E007406/1, EPSRC studentships with OMG VICON and Electrosonic Ltd, and made use of the octave facility funded by HEFCE. The authors would also like to thank the eye-catching team who are listed at: http://www.cve.salford.ac.uk/page/eye_catching

References

[1] J. C. Tang, E. A. Isaacs, and M. Rua, “Supporting distributed groups with a Montage of lightweight interactions,” Proc. ACM Conf. Computer Supported Cooperative Work (CSCW '94), October 22 - 26, 1994, pp.23-34.

[2] J. Hauber, H. Regenbrecht, M. Billinghurst, and A. Cockburn, “Spatiality in videoconferencing: trade-offs between efficiency and social presence,” Proc. ACM Conf. Computer Supported Cooperative Work (CSCW '06), November 04 - 08, 2006, pp.413-422.

[3] D. Roberts, R. Wolff, J. Rae, A. Steed, R. Aspin, M. McIntyre, A. Pena, O. Oyekoya and W. and Steptoe, “Communicating Eye-gaze Across a Distance: Comparing an Eye-gaze enabled Immersive Collaborative Virtual Environment, Aligned Video Conferencing, and Being Together”, Proc. IEEE Conf. Virtual Reality 2009, (IEEE VR 2009), March 14-18, 2009, IEEE Computer Society, pp.135- 142.

[4] S. Purdy, D. Barnhart, R. Johnston, R. Wuerfel and R. Ewart, “Latency Measurements Obtained from the Simulation Network Analysis Project,” Proc. Second int. workshop on Distributed interactive Simulation and Real-Time Applications (DIS- RT '98), July 19 - 20, 1998, IEEE Computer Society, pp.71.

[5] R. Hubbold, R. van Liere, and S.R. Ellis ”Latency in Virtual Environment Systems,” Proc. of IEEE Conf. Virtual Reality 2009 (IEEE VR 2009), March 14-18, 2009. IEEE Computer Society, pp.312-313.

[6] J.R. Kim, I.K. Park, and K.H. Shim, “The Effects of Network Loads and Latency in Multiplayer Online Games,” Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 4740/2007, pp.427-432.

[7] R. Wolff, D.J. Roberts, and O. Otto, “A study of event traffic during the shared manipulation of objects within a collaborative virtual environment”' Presence: Teleoperators and Virtual Environments. 13, 3 (Jul. 2004), pp.251-262.

[8] A., Murray, N., Rae, J., Steptoe, W., Steed, A., and Sharkey, “Communicating Eye Gaze across a Distance without Routing Participants to the Spot,” Proc. 11th iIEEE ACM int. symp. Distributed Simulation and Real Time Applications (DS-RT 08), IEEE Computer Society 2008, pp. 111-118, doi: 10.1109/DS-RT.2008.5

[9] D. He, F. Liu, D. Pape, G. Dawe and D. Sandin ”Video-Based Measurement of System Latency,” International Immersive Projection Technology Workshop (IPT 2000), 2000.

[10] A. Steed 2008, “A simple method for estimating the latency of interactive, real-time graphics simulations”' Proc. ACM symp Virtual Reality Software and Technology (VRST '08), ACM, New York, NY, pp.123-129.